Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Individually Contributed By:

Akash.K (akash.tup@gmail.com)

Please paste the GitHub Repo link.

Github Link:-
https://github.com/AkashKarthikeyan/Cardiovascular_Risk_Prediction_Capstone_Project

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Variables

Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

Application of SMOTE is carried out to oversample the minor class. SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.

Firstly, EDA is done with Standardisation, Feature Selection, and Encoding of Categorical Variables.

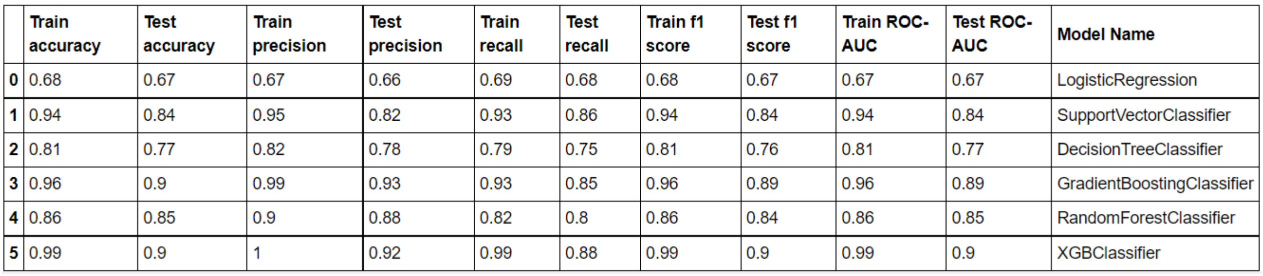
For modelling I tried various classification algorithms like:

1. **Logistic Regression**
2. **SVM Classifier**
3. **Decision Trees**
4. **Random Forest Classifier**

5. **Gradient Boosting Machine**
6. **XGBoost classifier**

Inwhich Hyperparameter tuning is carried out and the SHAP library is used to get the most important features. I have applied SHAP value plots on the Random Forest model to determine the features that were most important while model building and the features that didn't put much weight on the performance of our model.

The Classification Metrics are given below for all of the six models applied in the Cardiovascular Risk Prediction:



	Train accuracy	Test accuracy	Train precision	Test precision	Train recall	Test recall	Train f1 score	Test f1 score	Train ROC-AUC	Test ROC-AUC	Model Name
0	0.68	0.67	0.67	0.66	0.69	0.68	0.68	0.67	0.67	0.67	LogisticRegression
1	0.94	0.84	0.95	0.82	0.93	0.86	0.94	0.84	0.94	0.84	SupportVectorClassifier
2	0.81	0.77	0.82	0.78	0.79	0.75	0.81	0.76	0.81	0.77	DecisionTreeClassifier
3	0.96	0.9	0.99	0.93	0.93	0.85	0.96	0.89	0.96	0.89	GradientBoostingClassifier
4	0.86	0.85	0.9	0.88	0.82	0.8	0.86	0.84	0.86	0.85	RandomForestClassifier
5	0.99	0.9	1	0.92	0.99	0.88	0.99	0.9	0.99	0.9	XGBClassifier

The best performed is the XGBClassifier with an accuracy of 90% which is achieved by applying SMOTE to the Dataset as the Class 1 is very low(i.e., ~15% of the total dataset).

Therefore, either the Gradient Boosting Machine or the XGBoost can be used for the Risk Prediction which has the highest F1 Scores as well as the Accuracy Scores. (Note: GBM uses the loss function in the first degree whereas the XGBoost model uses the second degree loss function to estimate the dependent variable)